Claims

[c1] 1.A method for directing a request involving an expected load to one server out of a plurality of servers, comprising the steps of:
selecting a server;
determining whether said selected server has remaining capacity to handle said expected load; and
directing the request to said selected server, only if said server has remaining capacity to handle said expected load.

[c2] 2. The method of claim 1, further comprising the step of:
providing at least one token associated with each of the plurality of servers; and
wherein said step of selecting a server includes the step of selecting at least one token associated with said server.;

3. The method of claim 2, wherein a probability of selecting a token associated with said server differs from a probability of selecting a token associated with at least one other of the plurality of servers.

[c3] 4 The method of claim 2, wherein said step of providing at least one token includes the step of:,
providing a number of tokens associated with each of the plurality of servers, wherein said number is proportional to a the load limitation of each of said plurality of servers.

[c4] 5 The method of claim 4, further comprising the step of skewing the a probability of selection of a at least one token associated with said server, said skewed probability being disproportionate to said the number of tokens associated with said server.

[c5] 6. The method of claim 2, wherein said step of providing at least one token includes the step of:
providing a number of tokens associated with each of the plurality of servers, wherein said number is disproportionate to a load limitation of each of said plurality of servers and said number is at least partly based on a priority of each of the plurality of servers.

[c6] 7. The method of claim 1, further comprising the step of:
changing said remaining capacity to reflect said expected load if said request is directed to said server.

[c7] 8. The method of claim 1, further comprising the step of:
selecting another server if said server does not have remaining capacity to handle said expected load.

[c8] 9. The method of claim 8, wherein said other server is part of the same set.

[c9] 10.The method of claim 8, wherein said other server is part of a reserve set.

[c10] 11. The method of claim 1, further comprising the step of:
resetting said remaining capacity for each time frame.

[c11] 12. A system for allocating requests among servers, comprising:
a plurality of servers;
a first memory divided into entries, with at least one entry associated with each server and including an indication of said server;
a second memory divided into entries, with at least one entry associated with each server and including a representation of a remaining capacity of said server; and
a selector for selecting from among said entries of said first memory.

[c12] 13. The system of claim 12, further comprising:
at least one other set of at least one server to which requests can be allocated if there is no remaining capacity in any of said plurality of servers.

[c13] 14. A program storage device readable by machine, tangibly embodying a program of instructions executable by the machine to perform method steps for directing a request involving an expected load to one server out of a plurality of servers, comprising the steps of:
selecting a server;
determining whether said selected server has remaining capacity to handle said expected load; and
directing the request to said selected server, only if said server has remaining

capacity to handle said expected load.

[c14] 15. A computer program product comprising a computer useable medium having computer readable program code embodied therein for directing a request involving an expected load to one server out of a plurality of servers, the computer program product comprising:

computer readable program code for causing the computer to select a server;

computer readable program code for causing the computer to determine whether said selected server has remaining capacity to handle said expected load; and

computer readable program code for causing the computer to direct the request to said selected server, only if said server has remaining capacity to handle said expected load.